037230-1-A

Judgments of Probability and Utility for Decision-Making

CAMERON R. PETERSON

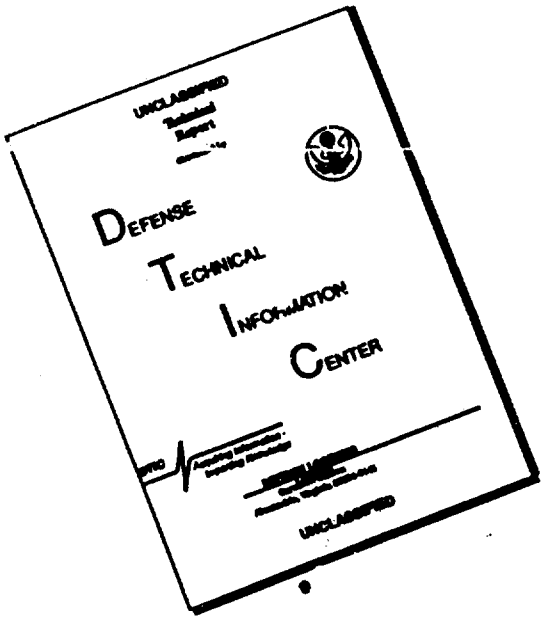
September 1971

The Department of the Navy
Office of Naval Research
Arlington, Virginia

Attention: Dr. Martin A. Tolcott
Director, Engineering Psychology Programs

Engineering Psychology Laboratory

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

JUDGMENTS OF PROBABILITY AND UTILITY FOR DECISION-MAKING

First Annual Report

30 September 1971

Contract Number N00014-67-A-0181-0034

NR 197-014

Cameron R. Peterson

Engineering Psychology Laboratory

The University of Michigan

Ann Arbor, Michigan

Prepared for:

The Department of the Navy
Office of Naval Research
Arlington, Virginia

Attention: Dr. Martin A. Tolcott
Director
Engineering Psychology Programs

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)

Department of Psychology
University of Michigan
Ann Arbor, Michigan 48105

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3 REPORT TITLE

MAN-MACHINE DECISION MAKING: JUDGMENTS OF PROBABILITY AND UTILITY FOR DECISION MAKING

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific - Annual Report for 12 months ending 30 September 1971

5 AUTHOR(S) (First name, middle initial, last name)

Cameron R. Peterson

6 REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
30 September 1971	41	6

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0181-0034	
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	none

10 DISTRIBUTION STATEMENT

1. This document has been approved for release and sale; its distribution is unlimited.

11 SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Engineering Psychology Programs Office of Naval Research

13 ABSTRACT

This ~~annual~~ report describes research intended to develop procedures for eliciting judgments of probability and utility that could be employed efficiently in a decision theoretic analysis. Experiments on probability estimation investigated (1) the reinforcing effects of a proper scoring rule upon probability estimates, (2) the use of Bayesian procedures to revise probability estimates, and (3) the relative merits of probabilities and odds as response modes. Research on the decomposition of utility estimates showed that (1) the model of a weighted linear average is relatively insensitive to nonadditive combination rules but highly sensitive to the nonlinearity of utility functions, (2) when utilities are esitmated it makes relatively little difference whether or not the judgments are decomposed, and (3) the procedures of decomposing utility estimates were feasible for use on a real world problem where water-quality engineers rather than college students served as subjects. Finally, a study is described which illustrated how the probability-revision procedures work on an actual naval problem--submarine surveillance.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

A-31408

TABLE OF CONTENTS

Security Classification

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
decision making						
probability						
utility						
Bayes theorem						
Proper scoring rules						
Multi attribute utility						

DD FORM 1 NOV 68 1473 (BACK)

S/N 0101-807-6821

Security Classification

A-31409

Introduction

This is a report of research supported by the Engineering Psychology Programs, Office of Naval Research, conducted under sponsorship of ONR Contract Number N00014-67-A-0181-0034, from the period of October 1, 1970 to September 30, 1971.

The purpose of the psychological research was to develop procedures for eliciting judgments of probability and utility that could be employed efficiently in decision theoretic analyses. Decision theory is a tool that is becoming increasingly popular in the field of business for making decisions about such things as building a new plant or introducing a new product. The basic approach is that a complicated decision is divided into its component parts. These components are evaluated and then arithmetic prescribed by decision theory is used to aggregate the evaluations into a final recommendation for a decision.

This approach offers promise for naval decisions as well as business decisions, but there is a serious stumbling block in that the evaluations for naval decisions are typically more complicated than for business decisions. The goal of naval decisions is not merely to maximize the expected rate of return; many non-monetary factors are important as well. Consequently, the evaluations which would serve as the input to a decision theoretic analysis of naval decision must be highly subjective. It is the goal of the research being conducted under the current contract to develop psychological procedures for eliciting subjective inputs to decision analyses.

The inputs to a decision analysis are to two kinds; they serve to answer the following question: "What are the stakes and what are the odds?"

The first kind of an input is called a utility. It is a number that measures the relative degree of attractiveness of a consequence of an action. The second measure is a probability, a number referring to the likelihood that the consequence will result if a decision is taken.

The first section of the report describes three sets of experiments on the problem of eliciting probability estimates; the second section describes a simulation and two experiments on the problem of eliciting utility judgments; and the third section describes attempts to test the feasibility of research on probability estimation in an operational naval environment, on the problem of submarine surveillance.

Probability

Nearly everyone understands that a probability is the number that refers to the likelihood of occurrence. The problem is that, with the exception of the forecasting of precipitation probabilities by weather forecasters, people have relatively little practice in attaching numerical estimates to their opinions about whether or not events will occur. Consequently, a major portion of the research conducted during the past year was aimed at finding procedures for eliciting probability estimates. This is necessary because such numerical judgments form one of the natural inputs to decision-making systems.

Validation of Probability Judgments

Research on how to elicit good judgments of probabilities presumes that the experimenter can identify a good judgment when he sees one. For probability judgments, researchers have used four criteria: optimality, accuracy, agreement with relative frequency, and consistency. The most stringent criterion, optimality, requires that the experimenter know the optimal probabilities. This is typically accomplished in experiments by using physical processes that produce equally likely elementary events, such as rolling fair die, to generate events. There is currently some debate about which scale to use for measuring the discrepancy between judged and optimal probabilities, but a log odds scale has desirable properties and is pretty well accepted for many situations.

The second criterion of a probability distribution, accuracy, can be measured by scoring rules, even without knowing the corresponding optimal probabilities, if it is possible to find out which of the alternative events turns out to be true. For example, a weather forecaster who estimates the probability of rain has no way of finding the optimal probability, but can find out whether or not it does

indeed rain. A high probability of rain is good if it does rain and a low one if it does not. Scoring rules essentially rate probability distributions as accurate to the extent that they pile up a lot of the probability on the event that turns out to be true. A scoring is said to be "proper" when the expected score can be maximized only by setting the judged probabilities equal to the corresponding optimal probabilities, when optimal probabilities are defined, or to observed relative frequencies, in situations in which it is possible to collect enough observations so that relative frequencies are good estimators of probabilities. It is therefore also true that the subjectively expected score can be maximized only by setting the judged probabilities equal to the corresponding subjective probabilities.

The third criterion requires that a distribution of judged probabilities agrees with the corresponding distribution of relative frequencies. This criterion is useful when a subject estimates probabilities of repeatable events, and the experimenter is able to measure the relative frequencies. It has been used extensively in experiments on learning, where there is interest in the degree to which estimated probabilities come to match the relative frequencies. They typically come very close indeed (Peterson & Beach, 1967). (Note that even in frequentistic situations this criterion is different from the accuracy criterion. The reasons why are subtle; it would lead the discussion too far afield to go into them here.)

A set of probabilities meets the fourth criterion, consistency, if it obeys the rules of probability theory. Optimal probabilities are consistent, but often not vice versa. Therefore, even though there is a strong tendency for probability judgments to be internally consistent (Beach, 1966; Peterson, Ulehla, Miller, Bourne, & Stilson, 1965), the criterion of consistency will be studied in the proposed research only as it relates to optimality or accuracy.

Use of Scoring Rules to Calibrate Probability Estimators

The use of quantitative judgmental inputs to a decision analysis requires that the person making the judgments be well-calibrated with respect to what the quantities mean. Thus, if a person is making a probability estimate he must have a good, intuitive appreciation of the probability or the odds scale upon which he is estimating. If he is making a value judgment he must understand the value scale that he is using. With respect to probabilities, research shows that some people are well-calibrated and others are not. Research shows, for example, across all days for which a weather forecaster has estimated, say, a 20% chance of rain, it turns out to rain nearly 20% of the time; and when he estimates a 60% chance of rain it turns out to rain approximately 60% of the time.

This quality of calibration, however, does not hold for college students. For example, I conducted an experiment several years ago in which each subject's task was to estimate whether or not a weak signal was contained in a noisy background. The subjects responded by estimating the probability that a signal was present in the noisy background. A smooth function could be used to describe the relation between percentage of trials containing signals and estimated probability of signal for each subject, but that function was not the identity line. When one subject estimated a 60% chance of signal it turned out that a signal was present about 80% of the time and when he estimated a 40% chance of signal, it was actually present only 20% of the time. That subject was conservative in his judgments about how much he had learned from his observations about whether or not a signal was contained in the noisy background. Across all trials on which he said that he had learned enough to do a 60:40 job of separating signals from noise, he had actually learned enough to do an 80:20 job. He had actually learned more about whether or not a signal was contained in the noise then he said he had learned. If such "hedging" were to occur in decision analyses of Naval

problems, the analyses would certainly turn out to be suboptimal.

Why are expert weather forecasters better calibrated than college students? There are probably several reasons, but one possibility is that weather forecasters have had considerable experience having their precipitation probabilities evaluated with a proper scoring rule. On days that it actually turns out to rain, a precipitation probability of 60% receives a higher score than does a precipitation probability of 40%. But how much better is the 60% estimate? Proper scoring rules have been developed to answer that question. A proper scoring rule has the property that a weather forecaster can maximize his expected score when estimating a precipitation probability only by estimating his true subjective probability. That is, if a weather forecaster expects that there is a 70% chance of rain then he will maximize his expected proper score only by estimating 70%. If a scoring rule is not proper it may be possible to receive a higher expected score by estimating a probability other than the true subjective probability.

Under the hypotheses that experience with a proper scoring rule was one of the factors that led to the better calibration on the part of weather forecasters, we designed the following experiment. College students were presented a long string of questions that were taken from encyclopedias, almanacs, and general knowledge examinations. An example of such a question is: "Is there a higher per capita income in Washington, D. C. or in the State of California?" The subject's task was to indicate which of the two answers he thought was more likely to be correct and then to estimate just how likely in terms of odds. For example, a subject may say that he thinks that California has the higher per capita income and that he estimates odds of 3:1 that he is correct.

There were two groups of subjects, experimental and control. During Stage One, the pretest, all subjects answered fifty questions like the one above

without receiving any kind of feedback about the right answer. Then, in Stage Two for the experimental group, subjects answered another group of seventy-five questions, and after each response they received feedback both in terms of which answer was right and also the proper score associated with the response given. The score was determined by the proper scoring rule designed for this experiment. In the control group subjects in Stage Two merely received feedback about which answer was correct; they were not informed that such a thing as a score existed. For Stage Three, all subjects took a post-test comprised of the same questions as the pre-test. That is, they once again gave odds estimates about questions asked in Stage One and once again they received no feedback.

The primary thing that distinguished the experimental group from the control group was that subjects in the experimental group received feedback that was generated by a proper scoring rule. Thus, if the use of experience with scoring rules serves to calibrate subjects with respect to odds estimates, then, from Stage One to Stage Three, odds made by subjects in the experimental group should improve more than the corresponding odds estimates made by subjects in the control group. Any greater improvement in the experimental group than in the control group should be attributable to recalibration by means of feedback generated through the use of scoring rules.

The results of the experiment support the experimental hypothesis. The proper scoring rule that was used in the training portion for the experimental group was also used to evaluate the quality of the odds estimates for both the pre-test given in Stage One and the post-test given in Stage Three. Of the twelve subjects in the control group, six improved from the pretest to the post-test and the scores for the other six became worse. This is essentially random performance, showing no more propensity to improve than not to improve. In the experimental

group, however, 10 of the 12 subjects improved and, on the average, the degree of improvement was rather substantial. Additional analyses of these results continue.

Encouraged by the results of this experiment, we administered the experimental procedure to 15 different intelligence analysts in different agencies in Washington D.C. where probability forecasts are beginning to be used. Results showed that the estimates for 12 of the 15 analysts improved from the pre-test to the post-test. Furthermore, the average amount of improvement was even more substantial than obtained for the college students. In addition to the quantitative results, several of the analysts volunteered comments to the effect that their intuitive "feelings" for the meaning of probability and odds changed as the result of the experience. Most analysts seemed to become more conservative in their estimates; they learned that it is costly to make estimates in the range of 50:1 without being very sure of the correct answer. On the other hand, at least one analyst who had a reputation of being cautious in probability estimates seems to have been making more extreme estimates since participating in the experiment. The scoring rule seems to be useful, at least for correcting extreme miscalibration.

We are conducting additional analyses in an attempt to understand the mechanism by which experience with a scoring rule improves calibration. When the analyses are complete we will write up the results and submit them to an appropriate scientific journal.

At the same time, we are beginning to develop a more streamlined version of the scoring rule test that can be self-administered to anyone who faces the task of estimating probabilities or odds. It can also be taken by people who are in a position to interpret probabilities estimated by someone else.

Bayesian Procedures to Revise Probabilities

In many practical situations, diagnosis about what the environment is like should precede deciding what to do about the environment. But unaided human intuition is subject to large and systematic biases in the process of a diagnosis. When a person estimates probabilities about which pair of hypotheses is true, and then revises his estimate in the light of new information, his revision is usually conservative. The revision is too small when compared with the optimal rule, Bayes's Theorem, for the revision of opinion.

This conservative bias obviously has important implications for any application of decision theory in which probability estimates must be updated. The biasing of inputs result in poor final decisions. Professor Ward Edwards at the University of Michigan has developed a set of procedures designed to reduce the conservative bias in updating opinion. He calls these procedures a probability information processing (PIP) system. It is based upon the assumption that people err in revising probability estimates because they misaggregate the impact of several data. A PIP system requires a person to estimate the diagnostic value of each individual datum (a likelihood ratio) and then assigns to a computer the task of aggregating across the individual likelihood ratios by means of Bayes's Theorem. Several experiments have shown that the degree of conservatism can be substantially reduced by the use of a PIP system.

The second experiment conducted during the last year of this contract was designed to investigate alternative procedures for using a PIP system. Some of our current on-line research with Naval officers is concerned with the problem of ship surveillance so it is appropriate that the general task faced by subjects in this experiment was one of trying to figure out the destination of a merchant ship. There were two possibilities. The ship was either going to Port A, which was relatively near and along a coastal route or the ship was going to Port B,

which was much further away and along an open-sea route. The experimenter displayed items of information upon which the subject was to base his judgment about the destination of the ship. These items of information included the amount of fuel taken on, the age of the ship in years, the size of the ship in terms of capacity for cargo and the percentage of capacity used on this trip. These data were related in obvious ways to the two ports and this relation was displayed to subjects by means of frequency distributions. For example, a display contained information about the amount of fuel that had been taken on by each of the last 100 ships that had sailed to Port A and it also contained the same kind of information about each of the last 100 ships that had sailed to Port B. On the average, of course, more fuel was taken on in the past by ships that were sailing to the more distant port.

Each subject made judgments about the destination of 26 different ships and he based his estimates on the four kinds of information presented in a counter-balanced sequence. That is, sometimes the subject learned about the age of the ship first and in other conditions he learned about amount of fuel first. The six different experimental conditions referred to six different procedures that were used to elicit judgments that could be used to calculate odds estimates about the destination of each ship. Those conditions are described below.

1. Unaided odds estimates.--Since this experiment employed judgments about ship destination, it was different from tasks used in previous experiments. So the goal of the first two conditions was to measure the degree to which results from which previous experiments (which typically involved abstract data generating processes such as the sampling of colored chips from a bag) could be generalized to the present task. The first condition employed direct odds estimates for which the subject simply wrote down which of the two ports he thought was the more likely destination of the ship; then he wrote down his revision of that odds estimate

after observing each of the four items of information. This is the kind of procedure that has typically produced conservative revisions and we anticipate that that will happen here.

2. Direct likelihood-ratio estimation.--A second condition that has been used frequently in past research on PIP systems requires the subject to estimate a likelihood-ratio upon observing each datum. Then after the experiment is complete the experimenter multiplies the likelihood-ratios together, as prescribed by Bayes's Theorem, in order to calculate the posterior odds with respect to ship destination. If the ship is rather new the likelihood-ratio should favor the distant, open-sea port and if the ship takes on only a small amount of fuel the likelihood-ratio should favor the near port. Previous research had shown that this procedure, using direct likelihood estimate as inputs to Bayes's Theorem, substantially reduces conservatism in the revision of odds estimates. That is the result we expect with this task.

3. Likelihood-ratio estimation with posterior odds feedback.--On-line research that we have recently conducted with intelligence analysts suggests that the direct estimation of likelihood-ratios as used in Condition Two will be unacceptable in practice. The reason is that it insulates the likelihood-ratio estimator from posterior odds. Therefore, after estimating a sequence of likelihood-ratios the analyst does not really know whether "system opinion" favors Port A or Port B but yet it is typically the analyst, who has specialized knowledge, who must defend that system opinion to his superiors and to operational officers. Accordingly, I doubt if the procedure of direct likelihood-ratio estimation will ever be used in practice, even though it may serve to reduce the conservative bias.

A natural extension of direct likelihood estimation is to display the posterior odds to the estimator. That is, each time the likelihood-ratio estimator makes an estimate, the implied posterior odds are calculated and then displayed. At least one

previous experiment conducted by Ward Edwards and his associates has shown that such posterior odds feedback cuts down on the efficiency of a PIP system in reducing conservatism. A serious difficulty with that experiment, however, was that there was no way of calculating the optimal posterior odds. It is therefore possible that the introduction of posterior odds feedback after the estimation of likelihood-ratios reduced the magnitude of posterior odds but not in the manner that was suboptimal. The current experiment will permit us to make that test. That is because the experimenter will control the process that generates the data about amount of fuel taken on, age of ship, and so on.

We expect the posterior odds resulting from the procedure in Condition Three will be less extreme than those of Condition Two, and we expect this "hedging" to be in a suboptimal direction, but the question is sufficiently important to warrant an experimental answer.

4. Likelihood-ratio estimation on a logrithmic scale with consistency checks.-- A potential benefit of using a Bayesian procedure is that the availability of individual likelihood-ratios permits the estimator to check the internal consistency of the likelihood-ratios. For example, after observing the sequence of data does the likelihood-ratio estimator really believe that the datum with the largest likelihood-ratio was the most diagnostic? And since Bayes's Theorem is multiplicative, because posterior odds is equal to the product of the likelihood-ratio and prior odds, the appropriate measure of the diagnostic value of a datum is equal to the log of the likelihood-ratio. For example, if the log of the likelihood-ratio for a datum is four times as great as the log of another likelihood-ratio, the first datum is four times as diagnostic as the second; it will take exactly four data like the second one to move the odds estimate as far as the single first datum.

The fourth condition was designed to exploit the possibility of consistency

checks. Does the procedure of requiring the likelihood-ratio estimator to make consistency checks among the logs of his likelihood-ratio estimates actually yield more optimal posterior odds?

These checks were made possible in the following manner. Within each sequence of data for Condition Four, the subjects drew four vertical lines on a logrithmic likelihood-ratio scale. After observing all four data, the likelihood-ratio estimator was asked to check the consistency among his estimates.

5. Odds and probability estimates on a log-odds scale.-- The fifth experimental condition was developed by the principal investigator and Mr.Clint Kelly while conducting on-line Bayesian experimentation with intelligence analysts in Washington, D. C. Several events combined to indicate that the use of straight likelihood estimates would be unacceptable in practice. The first was mentioned above; it will in general be necessary for the expert who estimates likelihood-ratios to know what the posterior odds are in order to defend those odds. In addition, although analysts find it reasonable to estimate likelihood-ratios under some conditions, they find it difficult or awkward in other conditions. For example, when the hypotheses about which the odds are being estimated have an apparent causal effect upon the datum observed, analysts find it more reasonable to estimate likelihood-ratios then when the reverse is true, then when the data seem to cause the hypotheses. For example, assume that an analyst is interested in whether or not King Heussin will remain in power for another year and the relevant datum is that there is an uprising by the fedayeen. The datum, the uprising, could have the ultimate effect of toppling Heussin from power and that is the kind of situation in which analysts find it difficult to estimate likelihood-ratios. Finally, analysts seem to find it more difficult to estimate likelihood-ratios when the datum is the nonoccurrence of an event rather than its occurrence. It is apparently psychologically easier to reason through plausible chains of events

that could result in the occurrence of an event rather than a change that could result in a nonoccurrence. It is this kind of difficulty with Bayesian procedures in on-line research that led the development of the procedure used in Condition Five.

This procedure requires that the analyst work with a combination of odds and likelihood-ratios. It is perhaps best understood by observing the response sheet on the following page, a response sheet that has been used extensively with intelligence analysts. The vertical axis of this response sheet indicates the odds on a logrithmic scale. The horizontal axis indicates the day of the month. The heavy line that traces a path through these coordinates is an example of how an analyst might estimate odds comparing the likelihoods of a pair of hypotheses as a function of events that occur on the indicated days. In this case, assume that an analyst's odds start out on the first of the month at 2:1 in favor of the H_1 (say, that a ship is sailing to the Atlantic) rather than H_2 (that it is sailing to the Mediterranean). Then on the 3rd of the month an event occurs that raises these odds to 4:1. On the 6th another event raises the odds to 8:1 and on the 7th of the month a third event drops the odds back to 2:1. This example yields the following interpretation: datum A and datum B have equal strength in that both are associated with likelihood-ratios of 2:1. Datum C, on the other hand, is as diagnostic as the combination of A and B. As an analyst moves his odds on this response form he may respond both in terms of odds and likelihood-ratios. The distance that he moves his odds as a result of a datum is a likelihood-ratio. When he attends to where he ends up rather than how far he is moving he is responding in odds. Intelligence analysts typically use both characteristics of this scale when responding. They attend both to how far they move as a result of the datum and to the resulting posterior odds.

Furthermore, after having processed several data it is possible for the analyst

to modify his estimates as a result of a consistency analysis. That is, he can examine the relative distances moved for each of several data. This scale permits the analyst to make a consistency analysis by examining the relative distances moved for the different data items and then adjust these movements for inconsistencies in case he thinks that one datum was more important but he originally moved with another datum further. If a reanalysis indicates to him that his original posterior odds were either too high or too low, it is possible to modify that estimate, but only by making corresponding revisions in movements along the way.

The illustrated form served to structure the response procedure for condition 5. The subjects used essentially the same form except that the horizontal axis referred to the datum number within a sequence rather than to the day of the month. Each sequence, of course, contained four data. Previous experience already indicated that analysts find this form acceptable for use. The purpose of the experiment was to evaluate the degree to which the use of this form could eliminate the conservatism which results from verbal odds estimates. A hypothesis is that condition 5 will result in less conservatism than condition 1 because the use of this chart requires the odds estimator to consider the impact of each data item individually. To a degree, the use of this chart reduces some of the requirements for intuitive aggregation inherent in direct odds estimates as used in condition 1.

6. The log-odds chart with probabilities deleted.-- Note that the right hand side of the chart in Figure 1 indicates the probability in favor of hypothesis 1 as implied by the corresponding odds that are displayed on the left side of the chart. However, the use of bounded probabilities as displayed in Figure 1, may inhibit movement toward either extreme. In order to test that hypothesis, subjects in condition 6 followed exactly the same

procedure as did subjects in condition 5, except that the probabilities were omitted from the response sheet for subjects in condition 6. The experimental hypothesis is that the deletion of the probability scale in condition 6 will lead to estimates that are more excessive than corresponding estimates in condition 5.

Data analysis.-- Subject running is now complete for this rather complicated experiment. A total of 40 subjects were run individually. Five subjects served in each of conditions 1, 2, 3, and 6, and 10 subjects served in each of the conditions 4 and 5 because those were the condition of major experimental interest. The subjects in the first three conditions participated for one to one and half hours, whereas subjects in conditions 4, 5, and 6 participated for three to four hours each. Greater subject-running time was required for the later conditions because of two factors. First of all, more training was required in order for subjects to understand the rationale behind the logrithmic response scale. In addition, subjects in the later three conditions were required to make internal consistency checks of the estimated or implied log likelihood ratios for each sequence of four data.

Analysis of the results has now begun. The major hypotheses about relative degree of conservatism will be tested by converting each subject's estimates to final posterior odds for each of the 26 sequences. These posterior odds will then be compared with optimal odds and also with corresponding posterior odds for other conditions. The effect of the consistency checks will be analyzed by inferring log likelihood ratios for each trial and each subject, and then correlating these log likelihood ratios with the corresponding optimal values. We hypothesize that the process of consistency checks will generate higher correlations, procedures to decrease conservatism will lead to higher regression slopes.

Probability Vs. Odds Estimates

Previous research has shown that subjects are more conservative when they revise probability estimates than when they revise odds estimates. That research however, used only two hypotheses, and an odds estimate, which is a ratio of two probabilities, seems naturally suited for only two hypotheses. It was important to learn, therefore, whether or not odds estimates will remain more effective with more than two hypotheses. That was the goal of the experiment to be described next.

The experiment made use of abstract stimuli; the subject's task was to infer the proportion of red chips in an urn that contained red chips and blue chips. The number of hypotheses were two, three and five, depending upon the experimental condition. With two hypotheses, the proportion of red chips could take on only two different values; with three hypotheses there were three possible proportions; and with five hypotheses the proportion of red chips could take on five different values. The task of the subject was to revise his estimate, either a probabiltiy or an odds estimate, on the basis of observing the color of chips sampled at random from the urn.

Because of large individual differences with this type of task, we used a within-subject design in which each of the twelve subjects participated in all three conditions. On some trials they made probability estimates and on others they made odds estimates. Probability revisions were made by redistributing 100 washers among different troughs where each trough represented one of the hypotheses. Odds estimates were made on a logrithmic scale; each individual estimate compared the probabilities for a pair of hypotheses.

Results.-- The results clearly demonstrated that advantage associated with odds estimates does not decrease as the number of hypotheses increase.

On the contrary, although odds estimates did not show much of an advantage with two or three hypotheses, they were substantially better than probability estimates when the number of hypotheses increased to five.

This experiment left us puzzled about why the odds estimates were not substantially more efficient than the probability estimates with only two hypotheses. But the experiment left no doubt about the fact that odds are substantially less conservative than probability estimates with several hypotheses, and this is the basic question that we had set out to answer. Our next step was to incorporate conditions 5 and 6 into the Bayesian experiment described in the preceding section in order to evaluate the relative merits of a joint probability-odds scale vs. odds alone. As indicated, the results of that experiment are now being analyzed.

Utility

One of the inputs to a decision analysis is a probability estimate and the other is an estimate of value or utility. The primary purpose of the decision-making system is to permit the choice of the best course of action, the one with the highest expected utility. Many approaches have investigated the utility or attractiveness of the consequence of an action. The simplest is to equate utility with dollars. Bernoulli understood problems with this approach as early as 1738 and argued that utility should be a function of money rather than simply equal to money. In this way he managed to capture the attitude of a decision maker toward risk as well as toward money. But many decisions are sufficiently complicated so that the possible outcomes of a course of action cannot be measured in terms of only money and risk. Frequently, other attributes such as time, effort, safety, and public opinion are also important. This problem has led to a serious interest in the study of multiattribute utilities. The basic problem is to discover the appropriate set of trade-off functions that will collapse several dimensions of value into a single dimension.

Validation of utility estimates.--There has been much less psychological research on eliciting utilities then on eliciting probabilities. The reason for the lack of research on utility is that it is difficult to validate estimates. A utility is usually considered to be a subjective quantity that characterizes the person making the judgment. Accordingly, a judgment of a utility is valid to the extent that it is close to the subjective quantity that it describes. The continuing difficulty in coming up with an independent measure of that subjective quantity has been the major stumbling-block to empirical research.

But there are ways around this problem. Our approach so far has been to postpone validation as long as possible. Suppose, for example, that we are interested in two different procedures for eliciting utility judgments. We wish to know which would be the best one to use in an applied setting. So far, we have adopted the policy of comparing the output of both procedures. If both outputs are the same, then a decision about which procedure to use can be based upon such practical considerations as convenience. Only when the two procedures differ with respect to output is it important to find which of the two is better. Accordingly, our present strategy has been to map out the kinds of situations in which different procedures yield similar results and the situations in which they yield different results; only when a difference exists will it be necessary to find the independent measure of utility in order to learn which procedure is best.

As will be shown below, our primary emphasis so far is to compare procedures that decompose the utility problem into a dimensional analysis with procedures that require the subject to aggregate across the dimensions intuitively. We have been finding a surprising degree of agreement between the output of the two procedures.

When we finally get to the stage of validation, we plan to use the concept of an organizational utility. The utility need not be restricted to the person making the judgment. In applications of decision-making systems, it is often the case that decisions should be made in such a way that they maximize the expected utility accruing to an <u>organization</u>, rather than the expected utility accruing to any individual within the organization.

This conception of a utility as a property of the organization makes it external to the person estimating it, and therefore opens the path to validation. The experimental procedure will be to create an imaginary organization with an externally specified multidimensional value system to train subjects about that value system, and then to invite the subjects to estimate utilities for multidimensional stimuli. The composite utility estimates can then be evaluated by comparing them with corresponding values for the organization.

Robustness of a Linear Model--A Simulation

A weighted linear average is the approach most frequently proposed for decomposing multidimensional utilities. With this approach, values along each dimension are multiplied by the relative weight of the dimension and then the products are added for a compositive utility. This approach requires the estimation only of weights for dimensions, not the utility functions. This simple linear model may not mirror the corresponding value system, but linear models frequently can account for much of the variance in systems that are largely non-linear. For example, Yntema and Torgerson in 1961 showed that the additive model could account for about 94% of the variance in a system with a non-additive, multiplicative combination rule.

Before completing the design of any experiment on the machine aggregation of utilities, however, we conducted further investigation of the degree to which different kinds of non-linear systems can be described by a linear model. In order to make this test we created four different kinds of two-dimensional value systems. One was linear and additive; so the linear additive model fit perfectly. A second

was linear and multiplicative; it was similar to the one used by Yntema and Torgerson. The third was non-linear but additive and the fourth was non-linear and multiplicative.

With each type of value system, we used two variables, each of which took on twenty values, and then measured the linear, multiple correlation for the set of all four hundred possible pairs of these two variables. The square of the multiple correlation denotes proportion of the variance that can be explained by the linear, additive model.

For the two cases with non-linear value systems, the relations between value and the underlying dimension were all monotonic. They included logrithmic exponential, and power functions (of the S-curved type). In each case we started with functions that were nearly linear and progressed toward functions that were highly non-linear.

Table 1 shows the results of this analysis. The columns refer to the four different forms of value systems, and rows refer to the degree

Table 1

Goodness-of-fit between linear additive models (measured by multiple R^2) and various value systems for the two variable case

Type of Data Generator

Case Number	1. Linear-Additive	2. Linear-Multiplicative	3. Nonlinear Additive	4. Nonlinear Multiplicative
1.	1.0	.87	.93	.80
2.	1.0	.92	.88	.80
3.	1.0	.89	.65	.52
4.	1.0	.95	.67	.64
5.	1.0	.97	.19	.07

increasingly non-linear ↓

of non-linearity for the non-linear systems. The first row is the most linear and the fifth row is the least linear. These results indicate, of course, that the linear, additive model accounts for all of the variance in the linear additive system. Furthermore, replicating the results of Yntema and Torgerson, the linear additive model accounts for nearly all of the variance in the linear, multiplicative system as shown in column two. It accounts for about 90% of the variance, on the average. But column three, showing the results of a non-linear, additive system, indicates that the amount of variance accounted for by the linear-additive model deteriorates markedly when the component value functions become nonlinear. Progressing from the first to the fifth row, the amount of the variance accounted for drops from 93% to 19%. Column four shows the same result. Predictability decreases markedly as the component value functions become nonlinear.

The implication of this analysis is simple and important. The linear, additive combination rule can be applied successfully to a value system only when the component value functions are highly linear. It makes little difference whether the combination rule is additive or multiplicative, but it makes a great deal of difference whether the component functions are linear or nonlinear.

This result is surprising in light of the degree to which the linear, additive model has been advertised as robust; and it was strong enough to deter us from using the linear additive approach for decomposing utility judgments in the experiments to be described below.

Experiment Decomposing Multidimensional Utilities

The purpose of this first laboratory experiment was to elicit utilities for multidimensional stimuli by using two different approaches - wholistic and decomposition - and then measure the agreement in resulting utilities from the two different procedures. To the degree that the outputs are alike, there is no need to worry about which approach is better. For the wholistic approach, subjects estimated the utility of a stimulus by a single number; for the decomposition approach, he estimated utility functions and weights for the component dimensions.

The purpose of the experiment required familiar stimuli with many value dimensions. After considering several kinds of stimuli, compact cars were chosen. The subject's task was to judge the relative utility of different cars. These cars varied along several dimensions, such as top speed in terms of miles per hour, economy as measured by miles per gallon, comfort as rated by expert judges, breaking performance, judged handling performance, and so on.

Two conditions differed in the amount of aggregation required. One condition used compact cars that differed only in three dimensions, whereas the condition requiring more aggregation used compact cars that differed in nine dimensions. The assumption was that any difference between the two procedures should increase with the amount of aggregation required.

The experiment employed three different response modes. The first two involved only direct estimation; a dollar scale in one case and a value scale ranging from 0 to 100 in the other. Thus, in the first condition the subject evaluated each car in terms of dollars (presumably a dimension with which he was familiar) and in the second case he evaluated the relative

values of the displayed compact cars on a 0 to 100 scale. A lottery was used for the third response mode because of an increasing interest in the concept of additive utilities. For example, would a subject gamble with a 50-50 chance of winning the most attractive car rather than the least attractive car; or would he prefer having a moderately attractive car for sure?

These, then, are the components of a three-dimensional factorial design: (1) decomposed vs. wholistic judgments; (2) 3 attributes vs. 9 attributes; and (3) the response modes of dollars, rating scales, and lotteries.

Result

Table 2 presents the correlational analysis of the results of the experiment. It displays correlations between the additive decomposed utility

Table 2

Correlations between decomposed and wholistic utility judgments

1. Three Dimensions	S#1	S#2	S#3	S#4	S#5	Median
a. Rating Scales	.99	.92	.95	.96	.94	.95
b. Dollars	1.00	.92	.97	.92	.92	.92
c. BRLTs	.91	.94	.94	.93	.93	.93
2. Nine Dimensions						
a. Rating Scales	.97	.96	.93	.91	.86	.93
b. Dollars	.98	.89	.97	1.00	.91	.97
c. BRLTs	.82	.85	.89	.84	.92	.85

models with the wholistic utility models. If the two approaches yield essentially the same utility functions, then the correlations should be about 1.0. In all cases the correlations are high; they are systematically below 90 in only one of the six conditions--when a lottery response mode was used with nine dimensions.

Table 3 presents a similar pattern of results. It displays differences between wholistic and intuitive approaches rather than a correlation between the two. The entries in Table 3 are mean squared differences between the

Table 3

Mean squared difference between decomposed and wholistic utility judgments - all scores normalized from 0 - 100.

1. Three Dimensions	S#1	S#2	S#3	S#4	S#5	Median
a. Rating scales	39	143	41	86	128	91
b. Dollars	3	234	73	204	151	151
c. BRLTs	272	306	119	155	183	183
2. Nine Dimensions						
a. Rating scales	45	103	127	216	257	127
b. Dollars	32	159	60	6	129	60
c. BRLTs	697	345	204	400	151	345

decomposed utility judgments and the wholistic utility judgments. All scales were normalized on a 0 to 100 scale to make results comparable throughout. Once again, the poorest performance resulted from the use of lotteries with

nine dimensions; next poorest from lotteries with three dimensions.

Taken together, these results suggest that there is a high correlation between decomposed and wholistic approaches. The absolute disagreement between the two approaches is greatest when the lottery response mode is used. Our first impression was that the use of a lottery was less efficient because of the greater complexity of the response. Additional reflection, however, suggested an alternative explanation. The direct estimation procedures resulted essentially in riskless utility functions, because they were derived from situations in which the subject assumed for sure that he would receive the stimulus car. The lottery procedure, on the other hand, generated utility functions that incorporated the judges' attitude toward risk, because there was uncertainty about whether or not the imaginary car would be recieved. Therefore, in additon to the greater natural complexity of the lottery task, there was also more to aggregate--attitude toward risk as well as the utility of dimensions along which the cars differed.

This result has prompted us to reconsider the entire theoretical approach to aggregating multidimensional utilities by means of an additive model. In many real world situations to which decision analysis will be employed there is uncertainty about whether or not the objects whose utilities are estimated will be received. In such a case, it is important to incorporate the decision-makers attitude toward risk as well as his attitude toward the values of each of the value dimensions. But that does not imply that the risk attitude must be tapped when measuring each of the value dimensions. An alternative procedure is to use one of the direct estimation procedures, such as rating scales or dollars, in order to collapse the set of value dimensions into a single dimension. Then, once that simplification

has been achieved, it is possible to find a single utility function which incorporates attitude toward risk on that single dimension, using traditional procedures for eliciting "utility for money" functions. We will propose additional research aimed at disentangling some of these ideas in the near future.

Field Research on Multi Dimensional Utilities

The field research that was conducted in parallel with laboratory research on multidimensional utilities is now complete. It forms Mr. Michael O'Connor's PhD dissertation, and the final draft is now being prepared. The purpose of this field research was to evaluate the degree to which procedures developed in the laboratory would work in the real world. Whereas laboratory research typically uses college students as subjects, and the tasks are relatively abstract, it is important to learn the degree to which these same procedures will be applicable when used with professional people who have a deep understanding of the problem on which the analysis is being applied.

This field work involved the construction of a water quality index. The construction of such an index is a task that is ideally suited for using decomposition procedures designed to cope with multidimensional utilities. The problem is that water quality is currently evaluated by gathering samples and measuring amounts of such polluting variables as fecal coliforms, phosphates, and so on. The problem is that there is no physical model that describes the trade-off functions between these dimensions. How much of an increase in fecal coliforms is required to exactly offset some specified decrease in suspended solids? Assuming that the less polluted water has more utility, Mr. O'Connor set about the task of finding these trade-off functions by using procedures of multi-attribute utility.

Mr. O'Connor made field trips to several different water quality engineers for the purpose of eliciting judgments of utility functions. The first roblem was to select a set of criteria to be included in the index. Almost immediately this posed a serious problem. What was desired was an overall measure of quality, but it became apparent very quickly that a very relevant question was utility for whom, or for what purpose. For example, consider the attribute of water temperature. If water is to be used for bathing, then a higher temperature is generally considered a good thing, but if it is to be used for industrial cooling then a higher temperature is a bad thing. Therefore, before constructing the quality index, it was decided to measure quality for two quite different purposes: for public water supply and for fish and wild life. One purpose of selecting two very different purposes of the utility function was to learn the degree to which the specific purpose was important in determining the function.

Then a list of dimensions or attributes was selected for each of the two functions by eliciting judgments from water quality engineers about which dimensions they felt were most important for each of the two functions. After settling upon the lists of dimensions, each engineer generated utility functions in the following way: for each attribute, he first drew a function describing the manner in which quality of water changed as the amount of the attribute increased. These functions were usually monotonic, but typically curvilinear. In a few instances they were nonmonotonic; the function increased until the concentration of the attribute was at an optimum, and then it decreased as the attribute exceeded that optimum. After estimating a quality

function for each attribute, the engineers then estimated relative weights. They did this by assigning a weight of 100 to the attribute considered most important, assumed a weight of 0 for an irrelevant, completely unimportant attribute, and then estimated intermediate weights for all of the other attributes.

These inputs were then sufficent to construct a multidimensional utility function for each of the water quality engineers. The functions were, of course, different, and so a modified delphi procedure was used to resolve the differences. Reasonably good consensus was achieved by the process.

A final index has been achieved for each of the two purposes: for public water and for fish and wild life. It was evaluated in two ways. First, the 11 participating water quality engineers agreed on the face validity of the index. Then several different samples of water were rated by the index. It turned out that even when the engineers differed with respect to their final utility weights or utility functions, the resulting correlations between engineers across water samples remained high. However, there was a much lower correlation between the two indices. These results indicate that small differences among the judges are not critical, but that it is important to determine exactly what the utility function is to be used for before generating the function. General purpose utility functions may be misleading.

Submarine Surveillance

Although it was not possible to field test the laboratory research on multidimensional utilities in a Naval setting, this was possible for research on the estimation of probabilities. Toward the end of the current contract year, on-line research was begun on the problem of submarine surveillance. This research was conducted with Naval intelligence officers in OP-942U (CNO USW Flag-plot) who are responsible for the surveillance of Soviet submarines. The research was conducted by Dr. Cameron Peterson and Mr. Clint W. Kelly; it was supported in part by the IBM Corporation and in part by this contract.

Procedures developed in the laboratory research described above were field tested on the problem of submarine surveillance. First, the scoring rule test was administered to the intelligence officers for purposes of calibration. Then we employed several different approaches to decomposition intended to improve the forecasts. The following example is an actual case study that is described here in hypothetical terms. It illustrates several of the *decomposition* procedures that have been used in other cases. In each instance, the numbers are those that were estimated by Naval intelligence officers, but some of the substantive portions of the case have been modified or disguised for purposes of security.

The problem is as follows: a submarine has been sighted leaving the Mediterranean Sea. The analysts were attempting to infer whether or not it was a nuclear submarine; was it an SSN or an SS? The following example illustrates several procedures used for estimating this probability.

Consider the two hypotheses: the first hypothesis, (H_1), is that it is an SSN and the second hypothesis, (H_2), is that it is an SS. Some historical data are relevant to this question. Six similar submarines

have been observed previously; five were SSNs and one was an SS. The task is to add that historical information to the analyst's theoretical knowledge in order to arrive at a probability estimate about this particular submarine. This estimate was achieved through the use of second-order beta probabilities and is illustrated in Figure 1. The top graph refers to the percentage of SSNs

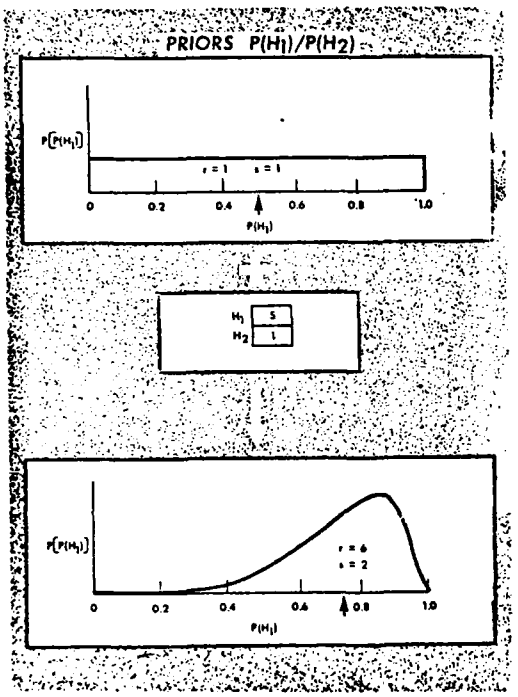


Figure 1

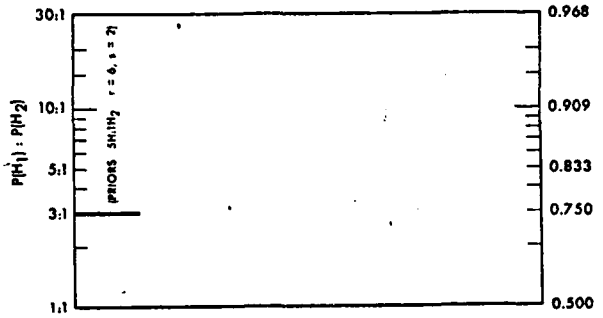


Figure 2

among all submarines that might be sent out. This is represented by $P(H_1)$ and is the horizontal axis of the top graph. The horizontal function is a rectangular probability distribution estimated by an intelligence analyst. This implies that, based upon theoretical knowledge and ignoring the historical frequencies, he expects that all proportions of SSNs are equally likely. It is

just as likely that they would send 15% as 40% as 80% SSNs. It is possible, through the use of beta functions, to combine this rectangular prior probability distribution with the historical frequency information in order to obtain a posterior probability distribution. The two parameters of the prior distribution, r=1 and s=1, are simply added to the relative frequencies in order to arrive at the appropriate parameters for the posterior probability distribution: r=6 and s=2. This posterior probability distribution is displayed in the bottom portion of Figure 1. It is a probability distribution over the proportion of SSNs. The mean or expectation of this probability distribution is .75 and so that is the number we selected as the prior probability that this particular submarine was an SSN. That is the number that served as a starting point for analyzing the following data.

Figure 2 shows the logrithmic chart that was used in the experiment on Bayesian procedures described above. The line at prior odds of 3:1 indicated on the left horizontal axis and the prior probability of .75 shown on the right horizontal axis indicates the starting point derived above.

The first datum is that another submarine was sighted on a homeward-bound course. For some weel-considered reasons, this datum slightly favors the hypothesis that the submarine being observed is an SSN; there is an estimated probability of .55 that this submarine would have been homeward-bound in about this time interval if the submarine being observed were an SSN; there is a 50% probability estimate if this were an SS. Therefore the likelihood ratio of the first datum, that a homeward bound submarine was observed, is .55/.50, which is equal to 1.1. This likelihood ratio is now inserted into the log odds chart as shown in Figure 3. It increases the odds in favor of an SSN by an almost imperceptible amount.

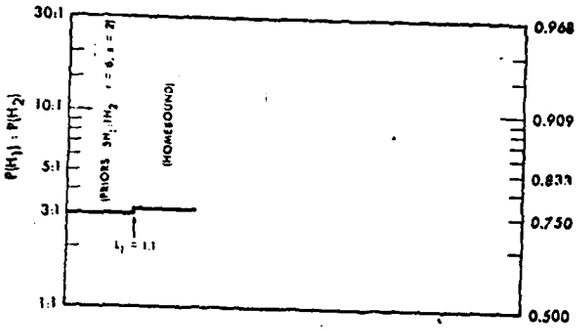


Figure 3

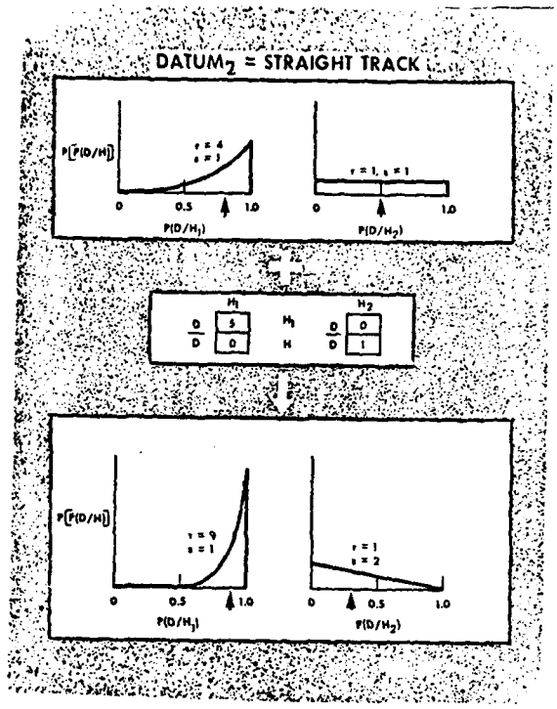


Figure 4

The second datum is that the submarine being observed seems to be following a straight track rather than taking evasive action. Figure 4 illustrates the manner in which the likelihood ratio associated with the straight track was elicited. This procedure also employed beta distributions. The upper portion of the figure refers to the prior beta distributions, ignoring historical frequencies of straight tracks. The left-hand function is the second order probability distribution over the proportion of straight tracks given an SSN as estimated by the analyst. The analyst expected that many more of the SSN, would follow straight tracks rather than evasive tracks.

The upper right-hand graph refers to the second order probability distribution for straight track given the second hypothesis, an SS. This is a uniform distribution. The middle portion of the figure shows the historical frequencies. All five SSNs previously observed had followed a straight track, whereas the one SS did not. Addition of these frequencies to the parameters of the beta distributions shown at the top of the figure imply the beta distributions shown at the bottom of the figure. The resulting likelihood ratio associated with datum 2 is therefore .90, the expectation of the bottom left-hand distribution, divided by .33. This likelihood ratio, 2.7, is now added to the log-odds chart as illustrated in Figure 5. The observation that the submarine is moving in a straight track has boosted the odds in favor of an SSN to approximately 9 to 1. This second datum is a very strong one, indeed.

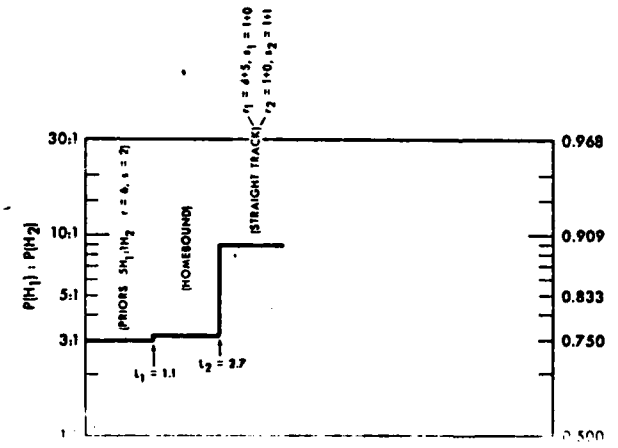


Figure 5

The analysis of the third datum is displayed in Figure 6. A political event that was expected to be observed was not observed. A conditional probability tree has been used to estimate the likelihood ratio. The left-hand branch of the tree refers to the two hypotheses, the SSN versus the SS. The next branch refers to the probability that the expected event

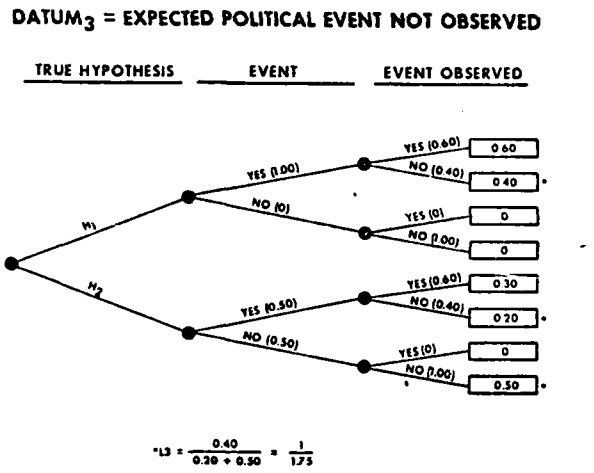


Figure 6.

would occur given each of the hypotheses. The analysts estimated that the event was certain to occur if they were observing an SSN, but the probability was only .50 that it would occur given an SS. The third column of branches refers to the probability of observing the event if it occurs. The analysts reviewed some experimental literature and concluded that there was a 60% chance that they would observe the event

if it actually occurred, but it was certain that they would not observe the event if it did not occur. That is, there was no chance of a false alarm. These branch probabilities implied the path probabilities displayed in the boxes at the right-hand side of the tree. Each probability was calculated by taking the product of the component branch probabilities. Thus, the .60 in the top branch is equal to 1.0 times .6. It is now a simple matter to find the likelihood ratio. The probability of not observing the expected event given as SSN is equal to .40 (the probability of observing it if it occurs, plus zero, the probability of observing it if it doesn't occur). The probability of not observing the expected event given an SS is equal to .70, so the resulting likelihood ratio is 1/1.75.

This likelihood ratio of 1/1.75 associated with not observing an expected event is now drawn on the log odds chart in Figure 7.

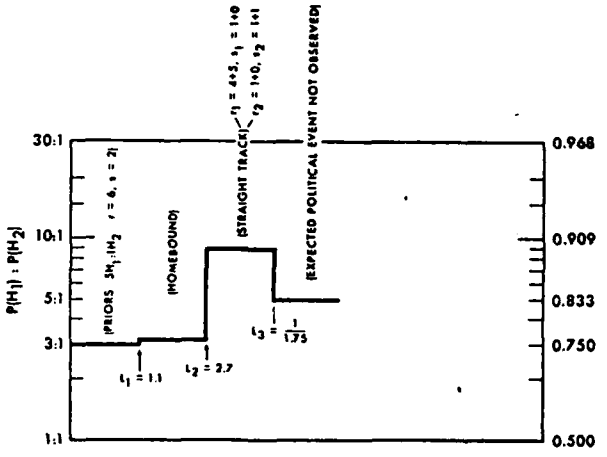


Figure 7.

It drops the odds in favor of an SSN to approximately 5 to 1.

The fourth datum is that there has been no contact with the submarine for an extended period of time. The analysts estimated a probability of 90% for an extended lack of contact with an SSN, but a probability of only 20% of no contact with the SS. Accordingly, the likelihood ratio associated with the fourth datum is .9/.2 or 4.5. This likelihood ratio is now added to the log odds chart as shown in Figure 8.

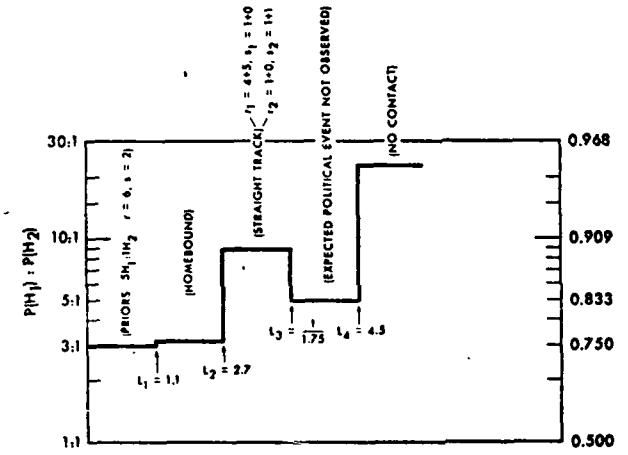


Figure 8.

It raises the odds in favor of the SSN to about 25 to 1.

The fifth datum later turned out to be a false alarm. An event has begun to develop and the analysts concluded that if it continued to develop it would favor the hypothesis that the submarine was an SS. The effect of this datum is displayed on the log odds chart in Figure 9.

Several attempts yielded no appropriate procedure for decomposing the estimate of the likelihood ratio associated with this particular event.

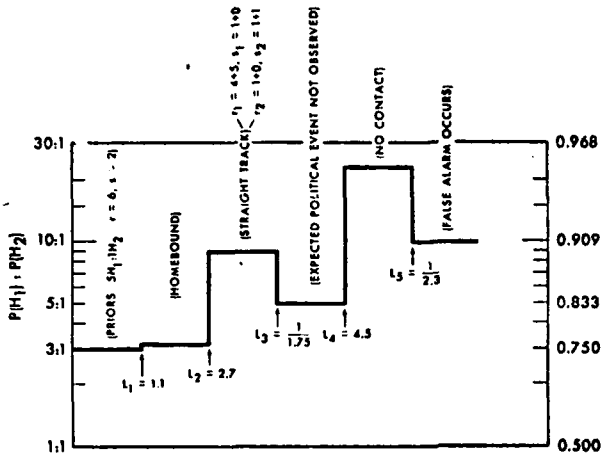


Figure 9.

Therefore, the analysts simply moved the odds on the log-odds chart on an intuitive basis. After considerable discussion, they decided that if the event did not turn out to be a false alarm, it favored the SS. It was somewhat more diagnostic than the datum of not observing the expected political event, and not quite as diagnostic as the straight track. They therefore moved the log-odds down just slightly less than they had moved it up as a function of the straight track. The resulting likelihood ratio turned out to be 2.3. The next morning, after completing this analysis, the final datum was identified as a false alarm and the odds estimates were therefore returned to about 25 to 1 as is shown in Figure 10. Some weeks later, after receiving much more information, it was

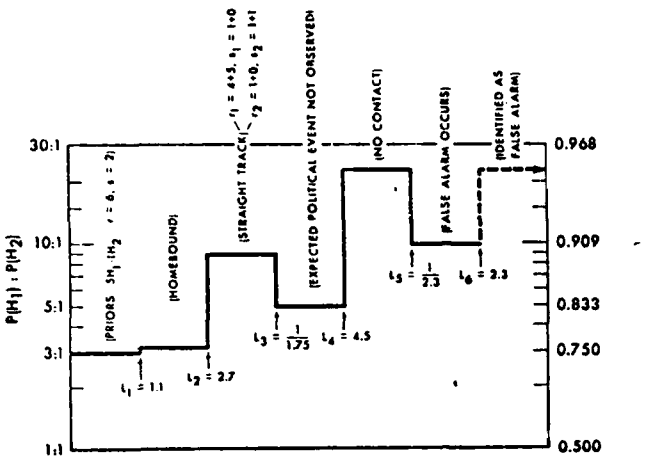


Figure 10.

concluded that this submarine was indeed an SSN. It is interesting to note that four of the analysts had discussed the problem thoroughly throughout the day and remained both uncertain and divided about which of the two hypotheses was correct. Yet the two analysts who made the estimates for this problem agreed with the final conclusion, that the data very strongly favored the hypothesis that they were observing an SSN.

We have now conducted several other case studies similar to the one described here. In general, the results have been encouraging with respect to the operational feasibility of techniques that are currently being developed for eliciting probability estimates and for revising those probability estimates in the light of new information. But these case studies have also highlighted some weaknesses of the laboratory procedures. Because of this usefulness, we intend to increase the field research with the intelligence analysts at OP 942U during the next contract year.

References

Beach, L. R., Accuracy and consistency in the revision of subjective probabilities. IEEE Trans. Hum. Fact. Electronics, 1966, 7, 29-37

Bernoulli, D. Specimen theoriae novae de mensura sortis. Comentarri Academiae Scientiarum Imperiales Petropolitanae, 1738, 5, 175-192 (Trans. by L. Sommer in Econometrica, 1954, 22, 23-36.)

Edwards, W., Phillips, L. D., Hays, W. L. & Goodman, B. Probabilistic Information Processing Systems: Design and Evaluation, IEEE Trans. System Sci. & Cybernetics, 1968, 3, 248-265.

Peterson, C. R. & Beach, L. R. Man as an intuitive statistician. Psychol. Bull. 1967, 68, 29-46.

Peterson, C. R., Ulehla, Z. J., Miller, A. J., & Bourne, L. E., Jr. Internal consistency of subjective probabilities. J. exp. Psychol. 1965, 70, 525-533

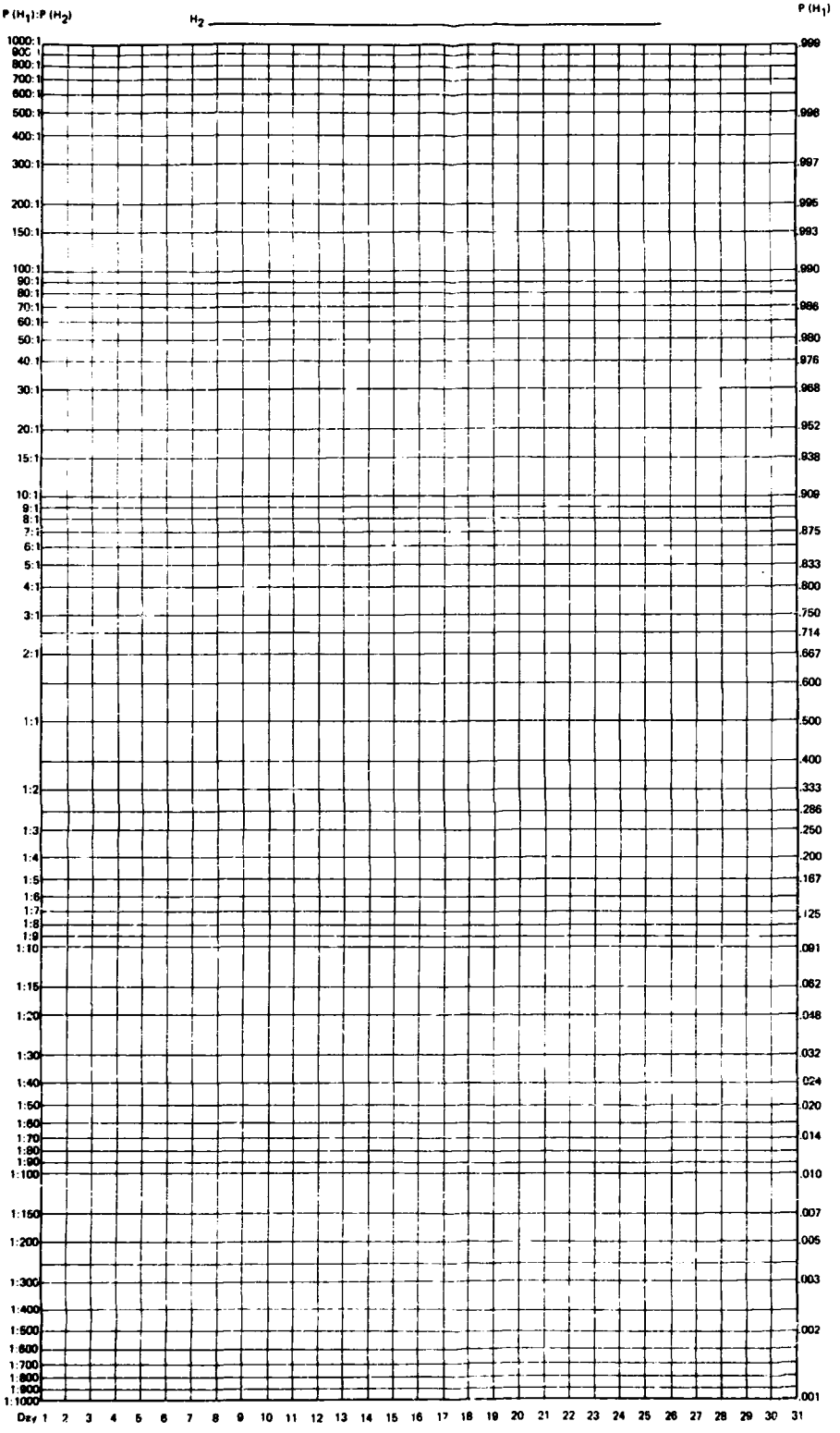
Yntema, D. B. & Torgerson, W. S. Man-computer cooperation in decisions requiring common sense. IRE Trans. Hum. Fact. in Electronics, 1961, 2, 20-26.

Personnel List

Cameron Peterson	Principal Investigator
Patricia Domas	Research Assistant
Gregory Fischer	Research Assistant
Jane Hoffman	Research Assistant
Michael O'Connor	Research Assistant
Linda Bender	Secretary
Patricia Homan	Secretary
Annette Johnson	Secretary

H_1 _____

H_2 _____


P (H1):P (H2)
P (H1)
1000:1
900:1
800:1
700:1
600:1
500:1
400:1
300:1
200:1
150:1
100:1
90:1
80:1
70:1
60:1
50:1
40:1
30:1
20:1
15:1
10:1
9:1
8:1
7:1
6:1
5:1
4:1
3:1
2:1
1:1
1:2
1:3
1:4
1:5
1:6
1:7
1:8
1:9
1:10
1:15
1:20
1:30
1:40
1:50
1:60
1:70
1:80
1:90
1:100
1:150
1:200
1:300
1:400
1:500
1:600
1:700
1:800
1:900
1:1000
.999
.998
.997
.995
.993
.990
.986
.980
.976
.968
.952
.938
.909
.875
.833
.800
.750
.714
.667
.600
.500
.400
.333
.286
.250
.200
.167
.125
.091
.082
.048
.032
.024
.020
.014
.010
.007
.005
.003
.002
.001
Day 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Month _____